INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# *What questions are my customers asking?*: Towards Actionable Insights from Customer Questions in Contact Center Calls

*Varun Nathan, Devashish Deshpande, Ayush Kumar, Cijo George, Jithendra Vepa*

Observe.AI, India

{varun.nathan, devashish.deshpande, ayush, cijo.george, jithendra}@observe.ai

## Abstract

Call centers serve as a critical point of contact between businesses and customers. The communication between customers and agents in such calls typically involves asking questions and responding to them. On average, a 10-minute call includes 2-3 customer questions. Such questions provide insights into customer's asks as well as identify areas of improvement for the questions where agents are taking longer time to respond. This motivates a need to peek into each call flowing through the contact center and derive business insights over such questions. To facilitate such deeper analysis and business intelligence at scale, there is a need to efficiently identify and rank the group of questions being asked over millions of calls flowing through a contact center. In this paper, we present a system for question monitoring via question extraction, rewriting and grouping, which enables contact centers to discover questions from calls at scale. Our in-house system leverages natural language processing techniques to transform customer questions into a format that is easily understandable, facilitating streamlined analysis of the data.

**Index Terms**: question extraction, question rewriting, question grouping, contact center AI, business intelligence

## 1. Introduction

The role of a contact center is to help an organization handle customer interactions via various channels like chat, email and phone calls. Of these, phone calls happen to be the most preferred method for customers to reach out to support teams owing to the convenience it offers in the form of live interaction with humans and faster resolutions. A major portion of these interactions in calls comprises questions coming from either party. Agents ask questions to comply with protocols they are expected to adhere to such as verification questions, questions for process adherence etc. In contrast, customers ask questions to find resolutions to issues they are facing with the products and/or services offered by the organization.

The questions asked by customers in call center conversations are of paramount importance as they provide critical insights into customer asks. By analyzing the types of questions being asked, call centers can gain a better understanding of customer behavior and refine their operations accordingly. For example, if a particular type of question is frequently asked, it may be an indication of a common issue that customers are facing, such as a problem with a product or service. In such cases, the call center can quickly identify the issue and work towards resolving it, thereby preventing further customer complaints and improving overall customer satisfaction. Furthermore, if agents are consistently receiving questions on a particular topic and are unable to provide satisfactory responses, it may indicate a knowledge gap or a need for better training in that area. By identifying such gaps, call centers can take proactive steps to address them, improving the overall quality of service provided to customers. In this paper, our focus is on detecting such customer questions with the motivation of driving actionable business insights as described in Section 3. In particular, we develop an end-to-end framework to:

1. Identify raw question spans and do a contextual question rewrite (Table 1) by resolving co-references and disfluencies to facilitate further analysis at scale.
2. Group customer inquiries for streamlined analysis along with associated call meta-data like time taken to respond to questions, thus driving informed decision-making.
3. Utilize a human-in-the-loop process to efficiently update the Knowledge Base with novel customer questions or optimize responses for questions that agents struggle to address, thus reducing response time and improving overall customer satisfaction.

Table 1: *Representative example of raw and re-written question.*

| |
|---|
| ***Agent***: good morning this XYZ financial services how may i assist you<br>***Customer***: okay i looking for the due amount of my account<br>***Agent***: sure let me check that for you i can see that the balance to be paid is two hundred and ninety dollars forty five cents we additionally charge for our service in form of service and sales tax<br>***Customer***: so now what alright how much would be the total like including that charge and all |
| Raw Question Span:<br>*how much would be the total like including that charge and all* |
| Contextual Rewritten Question:<br>*How much is the total due amount including service and sales tax?* |

## 2. Methodology and Pipeline Components

### 2.1. Automatic Speech Recognition

We ingest the raw audio calls in English-US language between agents and customers in our platform. To facilitate further steps of analyzing questions being asked, in the first step (Figure 1), we use a third party ASR Engine to get the transcript of the calls. The performance of ASR system evaluated across different verticals like, e-commerce, healthcare, financial services, retail etc. stands at 17.4 WER (word-error-rate).

### 2.2. Question Extractor

Upon obtaining the transcript of a call center conversation, the subsequent crucial step involves identifying the exact question spans within the dialogue. To facilitate this, we present a ma-

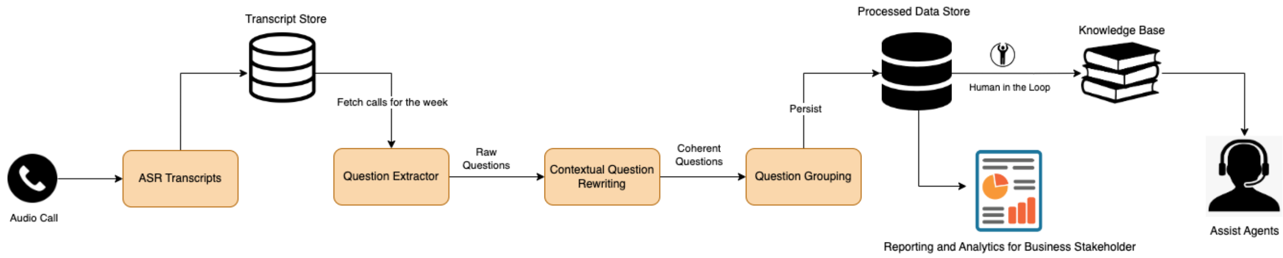


Figure 1: *Pipeline and components of the system to derive actionable insights over millions of contact center audio calls.*

chine learning model for detecting question spans, i.e, the beginning and end positions of a question within a given turn of the conversation. Our Question Span Detection (QSD) modelling is formulated as a sequence labelling problem on a dataset of 13.8k turn-question span pairs, achieving a promising Rouge-L score of 84.2 on the test set.

### 2.3. Contextual Question Rewriting

In call center conversations, disfluencies and contextual references can make it hard to accurately interpret customer queries until unless context is taken into account. To address this, we do a contextual question rewriting that uses contextual dependencies to produce a more concise and accurate representation of the customer's question (Table 1). Our approach involves training a T5-based [1] in-house large language model (LLM) that takes the question and context as input and generates a rewritten question that captures the intended meaning while filtering out irrelevant noise. Contextual question rewriting model obtains a Rouge-L score of 58.2 on a test set.

### 2.4. Question Grouping

Handling a high volume of customer inquiries can result in an overwhelming list of queries that are difficult to analyze one-by-one basis. To address this challenge, we use an unsupervised clustering algorithm that groups customer queries into coherent clusters. Our clustering pipeline uses BERTopic [2] which uses Sentence Transformers to embed questions into the vector space and then Hdbscan, a density based clustering algorithm, to cluster the embeddings. It is evaluated on the basis of topic coherence and topic diversity on the generated clusters.

## 3. Applications driven by Extracted Customer Questions

### 3.1. Reporting and Analytics for Business Stakeholder

High call volume in contact centers makes analyzing raw lists of customer questions impractical. Grouping questions provides insight into question categories and allows for deeper analysis as needed. Our Reporting and Analytics dashboard (Figure 3) includes metadata such as response time and agent information, allowing for different opportunities to be identified and acted upon, such as:

1. Frequent customer inquiries in a specific category can indicate underlying issues. For instance, if customers ask agents to repeat themselves often, it could indicate poor communication or audio quality. If many customers ask why the app is not available on Android, it may be an opportunity for the business to expand their base to Android users.
2. If an agent (or a group of agents) are taking longer time to respond to questions, they can be trained to reduce the operational costs by reducing average handle time (AHT). For example, agents might take longer to respond questions wrt a new product/campaign launch which indicates the need for a more thorough training to respond to questions related to the recently launched product/campaign.

### 3.2. Maintaining the Knowledge Base

In many contact center scenarios, agents rely on knowledge bases (KBs) to retrieve information pertaining to customer inquiries. The extracted questions and the categories of questions can be used to assesses whether any relevant information is stored in the KB. If not, the administrator may opt to add the new question and associated information to the KB, thereby ensuring that agents are equipped with the necessary knowledge to address similar customer queries in the future.

This approach has the potential to improve the overall quality of service provided by contact centers, by enabling a more efficient and effective handling of customer queries. Furthermore, the proposed system facilitates ongoing improvements to the KB, which can enhance the knowledge and expertise of the agents, and ultimately result in increased customer satisfaction.

## 4. Conclusion

We have presented a system for question extraction, rewriting, and grouping that enables to extract and analyze customer questions from millions of calls flowing through a contact center at scale. Our system provides several benefits for businesses. Firstly, by extracting and grouping customer questions, businesses can gain insights into the needs of their customers and identify areas of improvement. Additionally, our system can help identify limitations in knowledge bases, and allow businesses to easily add new information to these bases to better serve their customers in the future. Furthermore, the system can identify questions for which agents are taking longer to respond to questions. This provides an opportunity for businesses to reduce operational costs through targeted coaching of agents.

## 5. References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[2] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based TF-IDF procedure," *CoRR*, vol. abs/2203.05794, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.05794